Areal data

Applied Spatial Statistics

Areal data

- Point-referenced data are measurements taken at a specific location
- Areal data are summaries of regions/areas
- Example: COVID-19 mortality by county
- Health and economic data are often made available at the areal level to protect privacy
- Some variables are only interpretable at a regional level, e.g., soybean yield
- Modeling dependence between regions requires new methods

Areal data modeling

- Our analyses of point-referenced data hinged on defining the distance between locations
- There are infinitely-many possible locations so we needed models to be valid for any n sample locations
- In a sense, areal data are easier to deal with because there are a finite number of locations (e.g., $n = 50$ states)
- What's the distance between North and South Carolina?
- Is NC closer to SC or Virgina?

Defining spatial adjacencies

- Rather than distances, areal data models usually use adjacencies
- Let W_{ij} be the weight assigned to regions i and j
- Define $W_{ii} = 0$ and assume $W_{ij} = W_{ji}$
- The most common weight matrix is $W_{ij} = 1$ if regions i and j are adjacent and $W_{ij} = 0$ otherwise
- If $W_{ij} = 1$ then regions i and j said to be neighbors
- The weights can also be non-binary measures of distance between regions

Areal data notation

- The response variable in region i is Y_i and

$$\mathbf{Y} = (Y_1, ..., Y_n)^T$$

- The covariates in region i are

$$\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$$

  and the $n \times p$ covariate matrix is $\mathbf{X}$
- Let $\mathbf{W}$ be the $n \times n$ adjacency matrix with (i, j) element W_{ij}
- The number of regions that neighbor region i is

$$m_i = \sum_{j=1}^{n} W_{ij}$$

Measuring spatial autocorrelation

- The variogram plots spatial dependence by distance
- For areal data we measure correlation between neighbors
- Moran's I is a measure of spatial autocorrelation

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_i r_j W_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}} = \frac{\mathbf{r}^T \mathbf{W} \mathbf{r}}{\mathbf{1}^T \mathbf{W} \mathbf{1}}$$

where $r_i = (Y_i - \bar{Y})/s_y$ is the standardized response,
$\mathbf{r} = (r_1, ..., r_n)^T$ and $\mathbf{1} = (1, ..., 1)^T$

Testing for spatial autocorrelation

- If there is no autocorrelation the expected value of I is

$$\mathrm{E}(I) = -\frac{1}{n-1} \approx 0$$

- Large I suggests there is autocorrelation

- Moran's I can be used to test

  $\mathcal{H}_0$: no correlation
  $\mathcal{H}_1$: positive autocorrelation

Testing for spatial autocorrelation

Monte Carlo approximation to the p-value

1. Generate N datasets with independent and identically distributed Y_i

2. For each simulated dataset, compute Moran's I to approximate the sampling distribution under the null hypothesis

3. The p-value is approximated as proportion of the N Moran's I statistics that exceed the observed value

4. If the p-value is less than 0.05, reject the null hypothesis and conclude there is spatial autocorrelation

Moran's I versus Geary's C

- Geary C is an alternative measure of dependence

$$C = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(r_i - r_j)^2 W_{ij}}{2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}}$$

- Small C indicates strong autocorrelation
- This measures local variation as in the variogram
- A p-value for the test of autocorrelation can be computed as for Moran's I
- It is a good idea to compute both I and C